Cell, Vol. 115, 787–798, December 26, 2003, Copyright ©2003 by Cell Press

Prediction of Mammalian MicroRNA Targets

Benjamin P. Lewis,[1,4] I-hung Shih,[2,4]
Matthew W. Jones-Rhoades,[1,2] David P. Bartel,[1,2,*]
and Christopher B. Burge[1,*]
[1]Department of Biology
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139
[2]Whitehead Institute for Biomedical Research
9 Cambridge Center
Cambridge, Massachusetts 02142

Summary

MicroRNAs (miRNAs) can play important gene regulatory roles in nematodes, insects, and plants by base-pairing to mRNAs to specify posttranscriptional repression of these messages. However, the mRNAs regulated by vertebrate miRNAs are all unknown. Here we predict more than 400 regulatory target genes for the conserved vertebrate miRNAs by identifying mRNAs with conserved pairing to the 5′ region of the miRNA and evaluating the number and quality of these complementary sites. Rigorous tests using shuffled miRNA controls supported a majority of these predictions, with the fraction of false positives estimated at 31% for targets identified in human, mouse, and rat and 22% for targets identified in pufferfish as well as mammals. Eleven predicted targets (out of 15 tested) were supported experimentally using a HeLa cell reporter system. The predicted regulatory targets of mammalian miRNAs were enriched for genes involved in transcriptional regulation but also encompassed an unexpectedly broad range of other functions.

Introduction

MicroRNAs are endogenous ~22 nt RNAs that can play important gene regulatory roles by pairing to the messages of protein-coding genes to specify mRNA cleavage or repression of productive translation (Lai, 2003; Bartel, 2004). The first to be discovered were the *lin-4* and *let-7* miRNAs, which are components of the gene regulatory network that controls the timing of *C. elegans* larval development (Lee et al., 1993; Wightman et al., 1993; Moss et al., 1997; Reinhart et al., 2000; Abrahante et al., 2003; Lin et al., 2003). More recently discovered miRNA functions include the control of cell proliferation, cell death, and fat metabolism in flies (Brennecke et al., 2003; Xu et al., 2003) and the control of leaf and flower development in plants (Aukerman and Sakai, 2003; Chen, 2003; Emery et al., 2003; Palatnik et al., 2003).

MicroRNA genes are one of the more abundant classes of regulatory genes in animals, estimated to comprise between 0.5 and 1 percent of the predicted genes in worms, flies, and humans, raising the prospect that they could have many more regulatory functions than those uncovered to date (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001; Lai et al., 2003; Lim et al., 2003a, 2003b). The regulatory roles of the vertebrate miRNAs in particular remain unknown. The possibility that many mammalian miRNAs play important roles during development and other processes is supported by their tissue-specific or developmental stage-specific expression patterns as well as their evolutionary conservation, which is very strong within mammals and often extends to invertebrate homologs (Pasquinelli et al., 2000; Aravin et al., 2001; Lagos-Quintana et al., 2001, 2002, 2003; Lau et al., 2001; Lee and Ambros, 2001; Ambros et al., 2003b; Dostie et al., 2003; Houbaviy et al., 2003; Krichevsky et al., 2003; Lai et al., 2003; Lim et al., 2003a, 2003b; Moss and Tang, 2003). Indeed, miR-181, one of the many miRNAs conserved among vertebrates, is preferentially expressed in the B lymphocytes of mouse bone marrow, and the ectopic expression of this miRNA in hematopoietic stem/progenitor cells modulates blood cell development such that the proportion of B lymphocytes increases (Chen et al., 2003). However, regulatory targets have not been established or even confidently predicted for any of the vertebrate miRNAs, which has slowed progress toward understanding the functions of these tiny noncoding RNAs in humans and other vertebrates.

Finding regulatory targets is much easier for the plant miRNAs. In a systematic search for the targets of 13 *Arabidopsis* miRNA families, 49 unique targets were found with a signal-to-noise ratio exceeding 10:1, simply by looking for *Arabidopsis* messages with near-perfect complementarity to the miRNAs (Rhoades et al., 2002). Confidence in many of these predictions was bolstered by the observation that the complementarity is conserved among rice orthologs of the miRNAs and messages (Rhoades et al., 2002), and many of the 49 have since been confirmed experimentally (Llave et al., 2002; Emery et al., 2003; Kasschau et al., 2003; Tang et al., 2003). These predicted targets were greatly enriched in transcription factors involved in developmental patterning or stem cell maintenance and identity, suggesting that many plant miRNAs function during cellular differentiation to clear regulatory gene transcripts from daughter cell lineages, perhaps enabling more rapid differentiation without having to depend on regulatory genes having constitutively unstable messages (Rhoades et al., 2002). An analogous search for near-perfect pairing between the miRNAs and messages of *C. elegans* and *Drosophila* genes did not uncover more hits than would be expected by chance (Rhoades et al., 2002).

More sophisticated methods for predicting targets of insect miRNAs have recently been published (Stark et al., 2003) or submitted (Enright et al. http://genomebiology.com/2003/4/11/P8). The method of Stark et al. (2003) provides lists of candidate target genes that when used in combination with additional biological criteria, including functional relationships shared among predicted targets of individual miRNAs, led to validation of six targets for two *Drosophila* miRNAs (Stark et al., 2003). The cur-

*Correspondence: dbartel@wi.mit.edu (D.P.B.), cburge@mit.edu (C.B.B.)
[4]These authors contributed equally to this work.

rent *Drosophila* analyses do not include estimates of false positive rates, leaving open the question of the accuracy of these methods in cases where predicted targets of a miRNA do not have clear functional relatedness.

In the present study, we describe an approach that predicts hundreds of mammalian miRNA targets and provide computational and experimental evidence that most are authentic, allowing us to begin to explore fundamental questions about miRNA:target relationships in animals. Pairing to the 5′ portion of the miRNA, particularly nucleotides 2–8, appears to be most important for target recognition by vertebrate miRNAs. As seen previously for plant miRNAs, the predicted regulatory targets of mammalian miRNAs are enriched for genes involved in transcriptional regulation. In addition, the predicted mammalian regulatory targets encompass an unexpectedly broad range of other functions. Indeed, several lines of evidence imply that the targets identified in this initial analysis are only a fraction of the total, supporting the possibility that miRNAs regulate the expression of a large portion of the mammalian transcriptome.

Results and Discussion

An Algorithm for Predicting Vertebrate MicroRNA Targets

To identify the targets of vertebrate miRNAs, we developed an algorithm called TargetScan (the TargetScan software is available for download at http://genes.mit.edu/targetscan), which combines thermodynamics-based modeling of RNA:RNA duplex interactions with comparative sequence analysis to predict miRNA targets conserved across multiple genomes (Figure 1). Given an miRNA that is conserved in multiple organisms and a set of orthologous 3′ UTR sequences from these organisms, TargetScan (1) searches the UTRs in the first organism for segments of perfect Watson-Crick complementarity to bases 2–8 of the miRNA (numbered from the 5′ end)—we refer to this 7 nt segment of the miRNA as the "miRNA seed" and UTR heptamers with perfect Watson-Crick complementarity to the seed as "seed matches"; (2) extends each seed match with additional base pairs to the miRNA as far as possible in each direction, allowing G:U pairs, but stopping at mismatches; (3) optimizes basepairing of the remaining 3′ portion of the miRNA to the 35 bases of the UTR immediately 5′ of each seed match using the RNAfold program (Hofacker et al., 1994), thus extending each seed match to a longer "target site"; (4) assigns a folding free energy G to each such miRNA:target site interaction (ignoring initiation free energy) using RNAeval (Hofacker et al., 1994); (5) assigns a Z score to each UTR, defined as: $Z = \sum_{k=1}^{n} e^{-G_k/T}$, where n is the number of seed matches in the UTR, G_k is the free energy of the miRNA:target site interaction (kcal/mol) for the kth target site evaluated in the previous step, and T is a parameter described below (UTRs that have no seed match are assigned a Z score of 1.0); (6) sorts the UTRs in this organism by Z score and assigns a rank R_i to each; (7) repeats this process for the set of UTRs from each organism; and (8) predicts

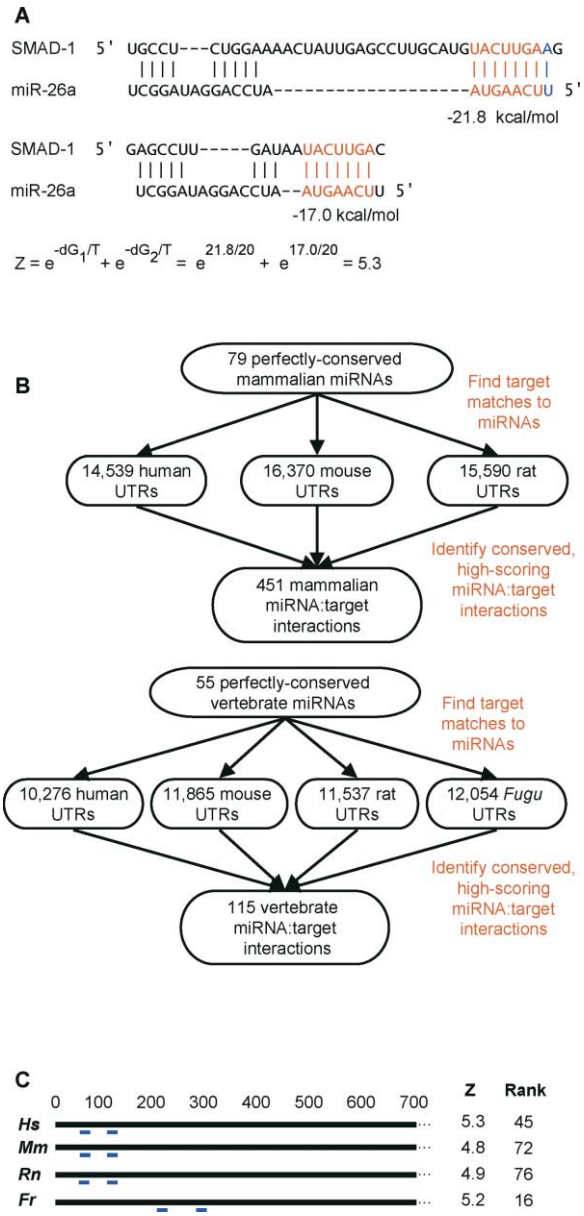

Figure 1. Prediction of miRNA Targets

(A) Structures, energies, and scoring for predicted RNA duplexes involving human miR-26a and two target sites in the 3′ UTR of the human *SMAD-1* gene, with seeds and seed matches in red and seed extension in blue.

(B) Schematic for identification of targets conserved across mammals (upper) and targets conserved in mammals and fish (lower). The number of genes from each organism with identified orthologs in every other organism is indicated.

(C) Positions of two target sites for miR-26a (blue) in orthologous *SMAD-1* 3′ UTR sequences from human (*Hs*), mouse (*Mm*), rat (*Rn*), and *Fugu* (*Fr*), with the Z score and rank of each miRNA:UTR pair, with $T = 20$.

as targets those genes for which both $Z_i \geq Z_C$ and $R_i \leq R_C$ for an orthologous UTR sequence in each organism, where Z_C and R_C are pre-chosen Z score and rank cutoffs.

The only free parameters in this protocol are R_C and Z_C, and the T parameter in the formula relating predicted

free energy to Z score. The value of the *T* parameter influences the relative weighting of UTRs with fewer high-affinity target sites to those with larger numbers of low-affinity target sites, and in this sense is analogous to temperature. However, there is no thermodynamic meaning to the *T* parameter or the Z scores used in this analysis; they merely provide a convenient means of weighting and summing predicted folding free energies. Suitable values for R_C, Z_C, and *T* were assigned by optimization over a range of reasonable values using separate training and test sets of miRNAs.

TargetScan was initially applied using two sets of miRNAs: a nonredundant pan-mammalian set of 79 miRNAs that have homologs in human, mouse, and pufferfish and identical sequence in human and mouse, but not necessarily pufferfish, and a nonredundant pan-vertebrate set of 55 miRNAs that have identical sequence in human, mouse, and pufferfish (Lagos-Quintana et al., 2001, 2002, 2003; Mourelatos et al., 2002; Dostie et al., 2003; Lim et al., 2003a). These sets, referred to as nrMamm and nrVert, respectively (Supplemental Table S1 at http://www.cell.com/cgi/content/full/115/7/787/DC1), are nonredundant in that when multiple miRNAs had identical seed heptamers, a single representative was chosen. The initial use of miRNAs that were both nonredundant and perfectly conserved among the queried species simplified the analysis of signal to noise.

Prediction of 400 Targets of Mammalian MicroRNAs at a Signal:Noise Ratio of 3.2:1

To predict mammalian miRNA targets, the nrMamm set of miRNAs was searched against orthologous human, mouse, and rat 3′ UTRs derived from the Ensembl classification of orthologous genes. Using $R_C = 200$, $Z_C = 4.5$, and $T = 20$, TargetScan identified 451 putative miRNA:target interactions (representing 400 distinct genes), an average of 5.7 targets per miRNA (Figure 2A). This number of predicted targets (the "signal") was compared to the number of targets predicted for cohorts of shuffled (i.e., randomly permuted) miRNAs (the "noise"). As described below, these shuffled sequences were carefully screened to ensure that our estimates of noise were as accurate as possible and not artifactually low. An average of only 1.8 targets were identified per shuffled miRNA sequence, for a signal:noise ratio of 3.2:1. This ratio was higher than the roughly 2:1 ratio observed for targets of the nrMamm miRNA set predicted using only the human and mouse UTRs (Figure 2A), underscoring the importance of evolutionary conservation across multiple genomes in our approach. The signal:noise ratio improved to 4.6:1 when conservation was required additionally in the fourth and most divergent species, *Fugu rubripes*, using the nrVert set of miRNAs (Figure 2A).

Although the signal:noise ratio improved as more genomes were included, the number of predicted targets per miRNA decreased—even though R_C and Z_C were relaxed to 350 and 4.5, respectively, and the value $T = 10$ was used for the four-species analysis (Figure 2A). Several factors might contribute to this effect, including the increased chance that an orthologous gene will be missing from the annotations of one genome as the number of organisms is increased. For example, the number of ortholog pairs available in human-mouse, 17166, decreased to 14539 ortholog sets in human-mouse-rat and 10276 ortholog sets in human-mouse-rat-*Fugu*. In addition, some miRNA:target interactions might not be conserved between mammals and fish. Another likely factor is that some features used by TargetScan to achieve an acceptable signal:noise ratio might not be strictly required for miRNA regulation. For example, although most known invertebrate miRNA target sites have 7 nt Watson-Crick seed matches (or longer matches), some do not, such as *lin-41,* a target of the *C. elegans* let-7 miRNA (Lee et al., 1993; Wightman et al., 1993; Moss et al., 1997; Reinhart et al., 2000; Abrahante et al., 2003; Brennecke et al., 2003; Lin et al., 2003). Thus, increasing the number of species increases the probability that the orthologous UTR of one or more species harbors functional sites that fail to satisfy the criteria required for TargetScan detection. Nonetheless, in 115 cases involving the UTRs of 107 genes, the predicted target sites were sufficiently conserved to be detected by TargetScan in orthologous UTRs from all four vertebrates (details of these predictions are given in Supplemental Table S5 and Figure S1A on the *Cell* website).

It is of utmost importance in this type of bioinformatic analysis to ensure that the shuffled control sequences preserve all relevant compositional features of the authentic miRNAs. For example, when compared to the seeds of shuffled cohorts that had not been screened to control for the expected number of target sites and the expected strength of miRNA:target site interactions, the seeds of vertebrate miRNAs have approximately 1.4 times as many seed matches in vertebrate UTRs. Specifically, the seeds of vertebrate miRNAs each had an average of about 2100 perfect-complement matches in masked vertebrate UTR regions whereas random heptamers with the same base composition averaged only about 1500 matches. The high number of additional matches seen for the miRNA seed (and also for the antisense of the seed) argues strongly against the biological significance of most of these matches. Instead, these excess matches appear to be the consequence of dinucleotide composition biases shared between vertebrate miRNAs and UTRs, which must be controlled for in order to avoid artificially high estimates of TargetScan signal:noise ratios (particularly in an algorithm that looks for multiple matches). Therefore, it was important to ensure that the shuffled miRNA controls matched the corresponding miRNAs closely in all sequence properties that impact the expected number and quality of TargetScan target sites. The properties we considered were (1) the expected frequency of seed matches in the UTR dataset; (2) the expected frequency of matching to the 3′ end of the miRNA; (3) the observed count of seed matches in the UTR dataset; and (4) the predicted free energy of a seed:seed match duplex. A miRNA shuffling protocol, MiRshuffle, was developed to generate randomized control sequences that possess all of these properties. For a given miRNA sequence, MiRshuffle generates a series of random permutations with the same length and base composition as the miRNA, until a shuffled sequence is found that matches the parent miRNA closely in each of the four criteria listed above.

The MiRshuffle procedure calculated expected frequencies using a first-order Markov model of 3′ UTR

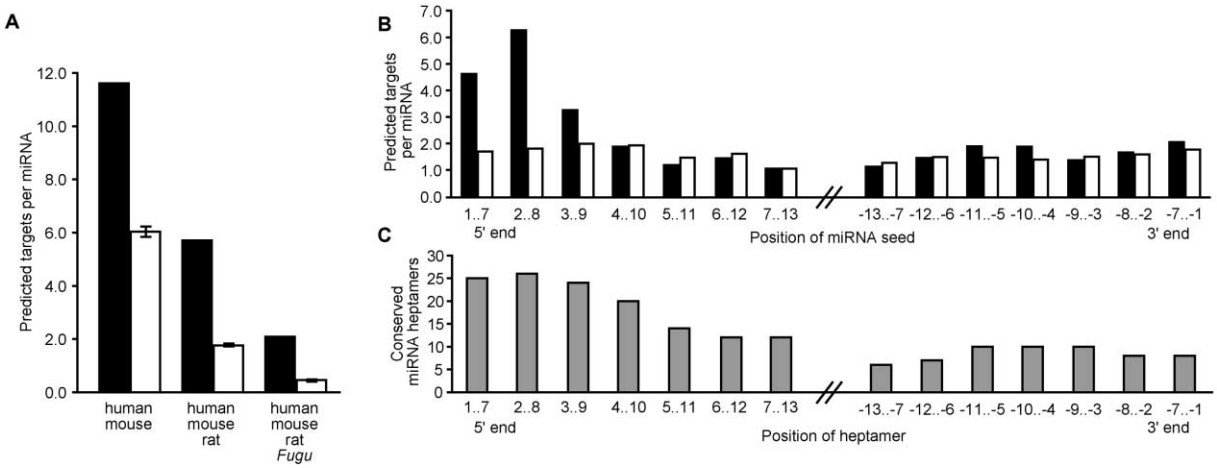

Figure 2. Predicted miRNA Targets Conserved in Multiple Genomes

(A) Mean number of predicted targets per miRNA for authentic miRNAs (filled bars) and mean and standard error of number of predicted targets per shuffled sequence for four cohorts of randomized miRNAs (open bars). Genomes used for identification of targets are listed below corresponding bars. The nrMamm set of 79 miRNAs was used for human/mouse and human/mouse/rat; the nrVert set of 55 miRNAs was used for human/mouse/rat/*Fugu*.

(B) Mean number of targets per miRNA using the human/mouse/rat UTR set and alternative miRNA seed positions for the nrVert miRNAs (filled bars) and for cohorts of shuffled controls (open bars). Positions of seed heptamer are indicated under bars; positive numbers indicate position relative to 5′ end of miRNA, negative numbers indicate positions relative to 3′ end of miRNA. Note that the signal:noise for the seed at 2..8 differs slightly from that of the human/mouse/rat analysis in (A) because a different set of miRNAs was used.

(C) Conserved heptamers among paralogous human miRNAs. For each position, the number of different heptamers that are perfectly conserved across multiple miRNAs in rMamm is shown.

composition that accounts for the long-recognized impact of dinucleotide frequency biases on the counts of longer oligonucleotides (Nussinov, 1981). As an additional control, another shuffling protocol was developed, DiMiRshuffle, which preserves the precise dinucleotide composition of both the seed and the 3′ end of the miRNA, as well as the seed match count and seed:seed match folding free energy. This protocol is less general than MiRshuffle in that not every oligonucleotide can be randomized while preserving exact dinucleotide composition—e.g., the only heptamer with the same dinucleotide composition as the miR-100 seed, ACCCGUA, is ACCCGUA itself. Nevertheless, it was possible to generate DiMiRShuffled controls for 47 of the 79 nrMamm miRNAs, and a signal:noise ratio of 3.5 was observed using this control in the three-mammal analysis (data not shown), comparable to the value obtained for MiRshuffled controls. Because of its wider applicability, MiRshuffle was used in all reported experiments.

In summary, even when the shuffled control sequences were carefully selected to closely match the corresponding miRNAs in all sequence properties expected to influence the number and quality of target sites, these shuffled controls yielded far fewer targets than did the authentic miRNA sequences. This difference results from an increased propensity of vertebrate UTRs to contain multiple conserved regions of complementarity to authentic miRNAs. We conclude that this propensity reflects a functional relationship between the miRNAs and the identified UTRs—that is, to the extent that the signal exceeds the noise, these identified UTRs are the regulatory targets of the miRNAs.

Correcting for the estimated rate of false positives, TargetScan appears to have identified an average of $5.7 - 1.8 = 3.9$ true targets conserved across mammals per miRNA (Figure 2A). A number of factors limit the sensitivity of our method, including (1) the incompleteness of orthologous gene annotations; (2) the possibility that some targets do not meet our stringent seed matching, Z score, or rank criteria; (3) the possibility that some mammalian target sites lie outside the 3′ UTR, as often observed for plant miRNAs (Rhoades et al., 2002); (4) the requirement that targets be conserved in the complete set of organisms; and (5) the limitation that our method does not model the simultaneous interaction of multiple miRNA species with the same UTR. Thus, the actual number of target genes regulated by each miRNA is likely to be substantially higher.

The Conserved 5′ Region of Mammalian MicroRNAs Is Most Important for Target Identification

TargetScan treats the 5′ and 3′ ends of miRNAs differently, with perfect basepairing required for the seed at the 5′ end, but no such requirement at the 3′ end. The importance of complementarity to the 5′ portion of invertebrate miRNAs has been suspected since the observation that complementary sites within the *lin-14* mRNA have "core elements" of complementarity to the 5′ segment of the *lin-4* miRNA (Wightman et al., 1993) and has been corroborated with the observation that the 5′ segments of numerous invertebrate miRNAs are perfectly complementary to 3′ UTR elements that mediate posttranscriptional regulation or are known miRNA targets (Lai, 2002; Stark et al., 2003). Moreover, the 5′ ends of related miRNAs tend to be better conserved than the 3′ ends (Lim et al., 2003b), further supporting the

hypothesis that these segments are most critical for mRNA recognition.

To explore this hypothesis, TargetScan was applied to predict targets of the nrVert miRNA set conserved between human, mouse, and rat using versions of the algorithm differing in the miRNA heptamer defined as the seed in step 1 (Figure 2B). Consistent with residues at the 5′ end of miRNAs being most important for target recognition, the highest signal:noise ratio was observed when the seed was positioned at or near the extreme 5′ end of the miRNA, with signal:noise values of 2.7, 3.4, and 1.6 observed for seeds at segments 1..7, 2..8, and 3..9, respectively, and signal:noise ratios of 1.3 or less at other seed positions. We suggest that the critical importance of pairing to segment 2..8 for target identification in silico reflects its importance for target recognition in vivo and speculate that this segment nucleates pairing between miRNAs and mRNAs.

Those seed positions that had the highest signal:noise ratios in the sliding seed analysis (Figure 2B) also had the highest degree of heptamer conservation in paralogous human miRNAs (Figure 2C). This observation strengthens the assertion that the signal seen above noise in our analysis reflects a functional relationship between the miRNAs and the identified UTRs because otherwise it would be difficult to explain why the most conserved portions of the miRNA and not other miRNA segments have the greatest propensity to match multiple conserved segments in UTRs.

The Number of Predicted Targets Is Greatest for the Most Highly Conserved MicroRNAs

The set of target genes predicted using conservation of miRNA complementarity across the three mammals was most suitable in size and quality for systematic analysis of gene function. To obtain as large a set of targets as possible, we searched our set of orthologous mammalian 3′ UTRs using an expanded set of 121 conserved mammalian miRNAs (rMamm, Supplemental Table S1 on *Cell* website) that includes miRNAs that were excluded from the nrMamm set because they had redundant seeds, yielding a total of 854 predicted miRNA:UTR pairs conserved across human, mouse, and rat (Supplemental Figure S1B). This number of predicted targets (854) represents an 89% increase over the 451 targets predicted for the nrMamm miRNAs, even though the number of miRNAs used increased by only 53% from 79 to 121. This discrepancy prompted us to ask whether membership in a multi-miRNA gene family influenced the abundance of targets. Indeed, we found that the 27 miRNAs in nrMamm that were members of paralogous miRNA families, i.e., families with variant miRNAs that have the same seed, had an average of 8.7 predicted targets per miRNA, more than twice the average of 4.2 seen for the remaining 52 nrMamm miRNAs, although the difference in signal:noise between these two sets was not as pronounced.

When initially expanding our list of mammalian miRNAs, we found that the set of 19 mammalian miRNAs that were conserved between human and rodents but for which a *Fugu* homolog was not found gave an unacceptably low signal:noise ratio of 1.2:1, even though the analysis did not extend to the *Fugu* UTRs. Accordingly, the rMamm set was restricted to those miRNAs with recognized *Fugu* homologs. The higher signal seen for the more broadly conserved miRNAs can be explained by the idea that miRNAs with larger numbers of targets would be under greater selective constraint, and therefore less likely to change during the course of evolution. Thus, more broadly conserved miRNAs would be likely to have more targets and consequently a higher TargetScan signal. This observation again supports the conclusion that TargetScan is detecting authentic targets because otherwise it would be difficult to explain the observed difference in signal:noise for broadly conserved miRNAs relative to that of less broadly conserved miRNAs.

The 854 miRNA:UTR pairs represented UTRs of just 442 distinct genes because many genes were hit by multiple miRNAs. In these cases, the miRNAs were usually, but not always, from the same paralogous miRNA family, often with the same seed heptamer. In those cases where the same UTR was hit by multiple miRNAs from different families (54 genes), the target sites generally did not overlap, consistent with simultaneous binding and regulation of some target genes by combinations of miRNAs. A complete list of the 442 target genes and the corresponding miRNAs is provided (Supplemental Figure S1B and Table S2 on the *Cell* website). An abbreviated list appears as Table 1, where genes were chosen on the basis of high biological interest. Genes involved in transcription, signal transduction, and cell-cell signaling dominate this list, including a number of human disease genes such as the tumor suppressor gene *PTEN* and the protooncogenes *E2F-1*, *N-MYC*, *C-KIT*, *FLI-1*, and *LIF*.

Experimental Support for 11 Predicted Regulatory Targets

Reporter assays were used to test 15 predicted targets of mammalian miRNAs in HeLa cells. The 15 targets selected for these experiments all had known biological functions but resembled the complete set of predictions in other respects, e.g., there was no significant difference in the average Z score, rank, or number of target sites per mRNA between the tested targets and the complete set of predicted targets. In only one case did the tested targets of a miRNA have obvious functional relatedness (NOTCH1, a receptor for DELTA1, both predicted targets of miR-34). Three of the 15 genes, *SMAD-1*, *BRN-3b*, and *Notch1*, were also in the set of predicted targets conserved to *Fugu*. Eight genes were predicted targets of miRNAs that had been cloned from HeLa cells (Lagos-Quintana et al., 2001; Mourelatos et al., 2002), and three genes were predicted targets of miR-34, which is also expressed in HeLa cells, based on Northern analysis (data not shown). For these 11 genes, a 100 to 1200 nt 3′ UTR segment that included miRNA target sites was inserted downstream of a firefly luciferase ORF, and luciferase activity was compared to that of an analogous reporter with point substitutions disrupting the target sites (as illustrated for *SMAD-1*, Figure 3A). Of these 11 UTRs, mutations in eight (*SMAD-1*, *SDF-1*, *BRN-3b*, *ENX-1*, *N-MYC*, *PTEN*, *Delta1*, and *Notch1*, but not *HOX-A5*, *MECP-2*, or *VAMP-2*) significantly enhanced expression ($p < 0.001$), as expected if

Table 1. Highly Cited Predicted Targets of Mammalian miRNAs

Category	Seed	miRNAs	Ensembl ID	Gene Name
Regulation of transcription/ DNA binding	AGUGCAA	miR-130,-130b	169057	Methyl-CPG-binding protein 2 (*MECP2*)
	GUGCAAA	miR-19a	169057	" "
	AAAGUGC	miR-20,-106	101412	Transcription factor *E2F1*
	GAGGUAG	*let-7*(a-g,i),miR-98	100823	DNA-(apurinic or apyrimidinic site) lyase (*APEN*)
	GAAAUGU	miR-203	125347	Interferon regulatory factor 1 (*IRF-1*).
	ACAGUAC	miR-101	134323	*N-MYC* protooncogene protein
	GAGGUAU	miR-202	134323	" "
	AAUCUCA	miR-216	065978	Nuclease sensitive element binding protein 1 (*YB-1*)
	UAAGGCA	miR-124a	163403	Microphtalmia-associated transcription factor
	GCUGGUG	miR-138	054598	Forkhead box protein C1 (*FKHL7*)
	AAAGUGC	miR-20,-106	103479	Retinoblastoma-like protein 2 (*RBR-2*)
	UCCAGUU	miR-145	151702	Friend leukemia integration 1 transcription factor (*FLI-1*)
	GCAGCAU	miR-103,-107	137309	High mobility group protein HMG-I/HMG-Y (*HMG-I(Y)*)
	GGAAGAC	miR-7	136826	Kruppel-like factor 4 (*EZF*)
Signal transduction/ cell-cell signaling	UAAGGCA	miR-124a	168610	Signal transducer and act. of transcription 3 (*STAT3*)
	UGGUCCC	miR-133,-133b	010610	T cell surface glycoprotein CD4 precursor
	UCACAUU	miR-23a,-23b	107562	Stromal cell-derived factor 1 precursor (*SDF-1*)
	GCUACAU	miR-221,-222	157404	Mast/stem cell growth factor receptor precursor (*C-KIT*)
	GGAAUGU	miR-1,-206	176697	Brain-derived neurotrophic factor precursor (*BDNF*)
	UAAGGCA	miR-124a	154188	Angiopoietin-1 precursor (*ANG-1*)
	GGCAGUG	miR-34	148400	Notch homolog protein 1 precursor (*HN1*)
	CCCUGAG	miR-125a,-125b	128342	Leukemia inhibitory factor precursor (*LIF*)
	AGUGCAA	miR-130,-130b	184371	Macrophage colony stimulating factor-1 precursor (*MCSF*)
	UCACAGU	miR-27a	184371	" "
	AAUACUG	miR-200b	008710	Polycystin 1 precursor
	GAAAUGU	miR-203	122641	Inhibin beta A chain precursor (*EDF*)
	AUUGCAC	miR-25,-92	065559	Dual spec. mitogen-activated protein kinase kinase 4
	GCUGGUG	miR-138	070886	Ephrin type-a receptor 8 precursor (*HEK3*)
	GUAAACA	miR-30(a-e)	156052	Guanine nucleotide-binding protein G(I),alpha-2 subunit
	AUUGCAC	miR-25,-92	156052	" "
	GAGAACU	miR-146	175104	TNF receptor-associated factor 6 (*TRAF6*)
	GGCUCAG	miR-24	166484	Mitogen-activated protein kinase 7 (*ERK4*)
	GAGAUGA	miR-143	166484	" "
	AGCUGCC	miR-22	166484	" "
	GCAGCAU	miR-103, -107	141433	Pituitary adenylate cyclase act. polypeptide precursor
Other	GUGCAAA	miR-19a,-19b	171862	Phosphatidylinositol-3,4,5-trisphos. 3-phosphatase (*PTEN*)
	AGUGCAA	miR-130,-130b	130164	Low-density lipoprotein receptor precursor (*LDLR*)
	GGAAUGU	miR-1,-206	160211	Glucose-6 phosphate 1-dehydrogenase (*G6PD*)
	UUGGCAC	miR-96	101986	Adrenoleukodystrophy protein (*ALDP*)
	AGCACCA	miR-29b,-29c	168542	Collagen alpha 1(III) chain precursor
	AGCACCA	miR-29b,-29c	114270	Collagen alpha 1(VII) chain precursor
	AUUGCAC	miR-25,-92	168090	*COP9* subunit 6
	AAGUGCU	miR-93	168090	" "
	AAAGUGC	miR-20,-106	168090	" "
	CCCUGAG	miR-125a,-125b	160613	Proprotein convertase subtilisin/kexin type 7 precursor

The 442 predicted targets conserved between human, mouse and rat were ranked based on the number of references listed in the RefSeq GenBank flatfiles (11/10/03 download). The top 37 most referenced predicted targets are shown, grouped on the basis of Gene Ontology annotations. The last six digits of the Ensembl ID are shown (ENSG00000#). MicroRNAs with different seeds that target the same UTR are listed on separate lines.

the endogenous miRNAs in the HeLa cells were specifying the repression of reporter gene expression by pairing to the predicted target sites (Figure 3B). Significantly enhanced expression was also observed when the analogous experiment was performed using either the full-length *C. elegans lin-41* 3′ UTR or a 124 nt segment of the UTR containing the two previously proposed *let-7* miRNA target sites (Reinhart et al., 2000), indicating that at least some of the repression of *lin-41* observed in *C. elegans* can be recapitulated by HeLa *let-7* miRNA in this heterologous reporter assay (Figure 3B). For all eight predicted human targets of endogenous HeLa miRNAs that responded to mutations, the increase in expression seen when disrupting the pairing to the miRNA seed was at least as high as that seen for mutations in the *let-7* target sites of *lin-41* (Figure 3B).

Four tested genes (*G6PD*, *BDNF*, *MCSF*, and *LDLR*) were predicted targets of miR-1 and miR-130, two miRNAs that had not been cloned from HeLa cells and were not detected by Northern analysis. Initially, reporters containing UTR segments from these four genes were examined for response to transfected miRNAs (Doench et al., 2003) (data not shown). Of the four, *G6PD*, *BDNF*, and *MCSF* responded to the transfected miRNAs. To further validate these targets, we used a second assay resembling the one described for targets of miRNAs expressed in HeLa cells, except that it took advantage of HeLa cell lines ectopically expressing either human miR-1 or human miR-130. Mutations in the miRNA target sites of all three of the genes that had responded to transfected miRNAs led to significantly increased reporter output in the lines expressing the cognate miRNAs, but not

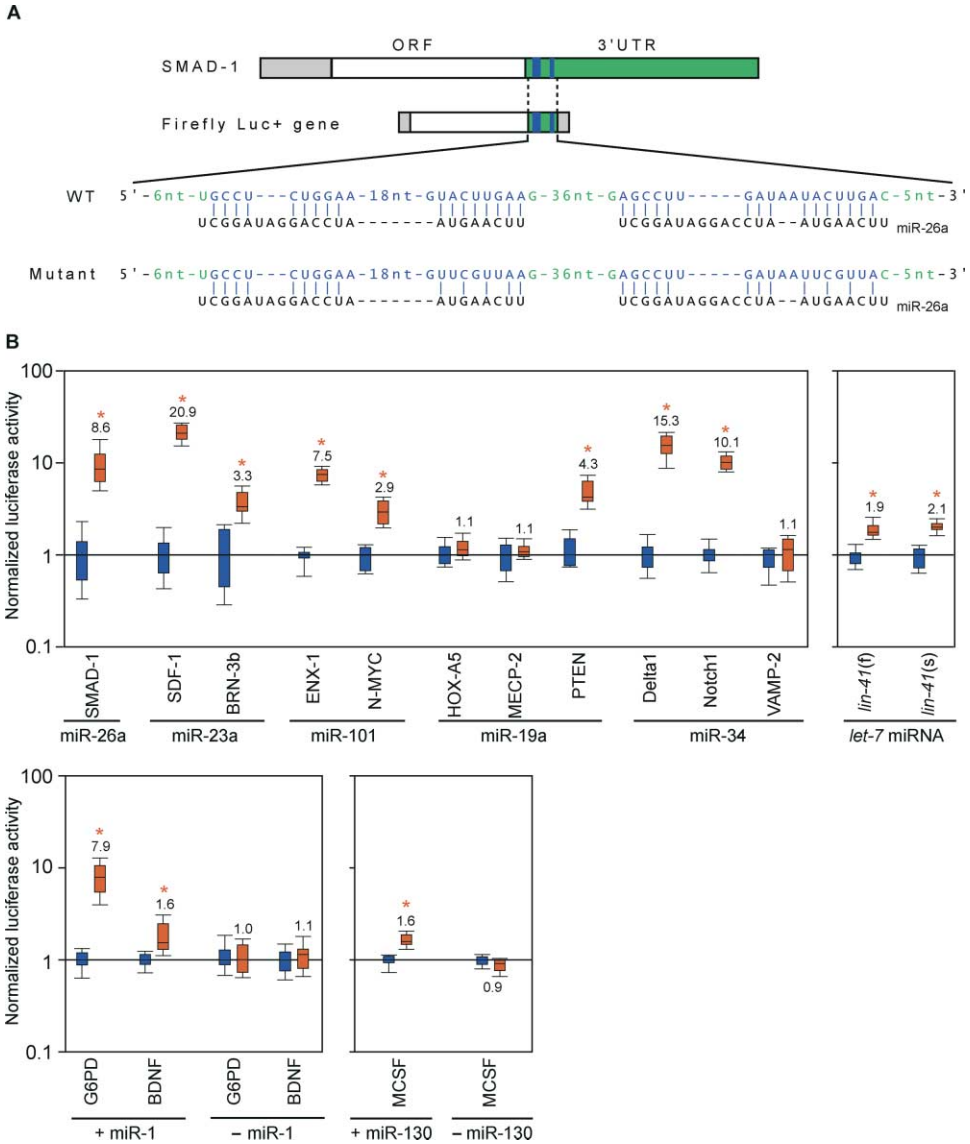

Figure 3. Experimental Support for Predicted Targets

(A) Schematic of a reporter construct used to evaluate the role of complementarity between miR-26a and the *SMAD-1* 3′ UTR. The wild-type (WT) construct had a 106 nt fragment of the *SMAD-1* UTR (green) containing two miR-26a target sites (blue) inserted within the firefly luciferase 3′ UTR. The mutant construct was identical to the WT construct except that it had three point substitutions (red) disrupting pairing to each miR-26a seed.

(B) Box plots showing the luciferase activity after reporter plasmids were transfected into HeLa cells. Reporters analogous to those depicted for *SMAD-1* were constructed for the indicated target genes (Supplemental Figure S2 on *Cell* website). The UTR fragments often had two target sites to the indicated miRNA, and both were disrupted in the mutant reporters (exceptions were *SDF-1*, *BRN-3b*, *G6PD*, *Delta1*, *Notch1*, and *BDNF*, which each had three target sites, two of which were disrupted, and *N-MYC*, which had one of its two miR-101 sites disrupted). Firefly luciferase activity was normalized to *Renilla* luciferase activity of the transfection control plasmid and then normalized to the median activity of the corresponding WT reporter. Each box represents the distribution of activity measured for each WT (blue) and mutant (red) reporter ($n = 12–15$; ends of the boxes define the 25th and 75th percentiles, a line indicates the median, bars define the 10th and 90th percentiles, and the number indicates the median activity of the mutant reporter). Asterisks (*) denote instances in which differences between the WT and mutant were statistically significant ($p < 0.001$; Mann-Whitney test). Two pairs of constructs for *C. elegans lin-41*, a previously known target of *let-7*, were tested, one with a full-length and the other with a 124 nt UTR segment (f and s, respectively). Except for miR-1 and miR-130, the miRNAs were all endogenously expressed in the HeLa cells. Reporters corresponding to predicted targets of miR-1 and miR-130 (*G6PD*, *BDNF*, and *MCSF*) were each examined in a HeLa cell line stably expressing the relevant miRNA (+ miR-1 or + miR-130) and the parental cell line (− miR-1 or − miR-130).

in the parental lines lacking the miRNAs (Figure 3B), as expected if these genes were authentic targets of the respective miRNAs. The levels of ectopically expressed miR-1 and miR-130 were comparable to those of endogenous miRNAs, as judged by Northern blot analysis (Lim et al., 2003b). For miR-1, Northern analysis with a synthetic miR-1 standard allowed accurate quantitation, revealing an average expression of 500 miR-1 molecules per cell.

In sum, for 11 of the 15 cases tested, the sites identi-

fied by TargetScan influenced expression of an upstream ORF when expressed in the same cells as the corresponding miRNAs. Additional experiments in animals will be needed to address the particular biological consequences of these regulatory interactions, but the evolutionary conservation of the pairings suggests that they are important. All four of the remaining genes might not be true targets; our statistical analysis using shuffled controls indicated that about 30% of predicted mammalian targets are likely to be false positives (Figure 2). Alternatively, some might still be authentic targets whose regulation was not detected in our assays. Regulation would be missed in cases for which cell type-specific factors were required that were not expressed in HeLa cells, or in cases for which additional mRNA elements were required but were not included in the UTR segments used in our reporters.

One limitation of the existing sequence databases that complicates the systematic identification of miRNA targets is that UTR annotations are often absent or incomplete. In order to compensate for this limitation, we had extended each annotated 3′ UTR with 2 kb of 3′ flanking sequence. Using extended UTRs substantially increased the number of predicted targets, with signal-to-noise ratios at least as high as they were for unextended UTRs, suggesting that extension of the annotated UTRs allows detection of many additional authentic target genes. One consequence of using this UTR-extension protocol is that for some genes, all predicted target sites will fall outside of annotated UTRs. Manual inspection of the 15 UTR regions tested in our reporter assays revealed that in all but one of these cases the tested target sites were contained within regions whose status as UTRs was supported by known ESTs and predicted polyadenylation sites, even though some of these regions are not yet annotated as human UTRs. For the single exception, the *Notch1* gene, the tested target sites were all located downstream of the annotated 3′ UTR of the human gene, and the end of the annotated *Notch1* 3′ UTR was supported by a predicted polyadenylation site and alignment of multiple ESTs. However, *Notch1* might have additional 3′ UTR isoforms; many human genes—perhaps as many as 50% or more of the genes in the genome—have alternative polyadenylation sites (Iseli et al., 2002). In order to investigate the potential expression of the tested *Notch1* target sites, which gave a positive result in our assay for miRNA regulation (Figure 3), an RT-PCR assay was used with polyA-selected RNA from a pool of human tissues. Consistent with the possibility that these sites lie within an alternative UTR isoform of *Notch1*, an RT-dependent product of the correct size and sequence was observed (data not shown). The TargetScan set of predicted mammalian target genes (Supplemental Table S1B on the *Cell* website) undoubtedly contains other examples for which the target sites all lie outside of the UTR regions supported by available data; some of these will be false positives, but others might point to the miRNA regulation of alternative mRNA isoforms.

Human miRNAs Predominantly Are Negative Regulators of Gene Expression

The finding that a sizable fraction of the tested UTR segments were sensitive to mutations disrupting their target sites supports the assertion that most of the predicted targets are authentic. For many, the pairing outside the seed was less extensive than that previously proposed for miRNA targets (Supplemental Figures S1A and S1B). Perhaps TargetScan is identifying mRNA elements that are necessary but not sufficient for miRNA regulation. Alternatively, these elements might be sufficient, in which case their low information content raises the possibility that miRNAs modulate the utilization of a substantial fraction of the mammalian mRNAs.

In none of the 15 cases tested was there evidence of miRNA-mediated activation of reporter expression; changes either were not statistically significant or were in the direction of miRNA-directed repression. This result suggests that mammalian miRNAs are generally negative regulators of gene expression, as has been observed for the known examples in invertebrates and plants (Lai, 2003; Bartel, 2004).

Predicted Mammalian MicroRNA Targets Have Diverse Functions

To assess target gene functions, we evaluated the frequency of specific gene ontology (GO) molecular function classifications (Gene Ontology Consortium, 2001) among the predicted targets of the nrMamm miRNAs and their shuffled control sequences (Table 2). Predicted miRNA targets populated many major GO functional categories, and for each of these categories, the number of targets for the real miRNAs greatly exceeded the average for the shuffled cohorts. Therefore, despite the presence of false positives among our predictions, the data in Table 2 strongly indicate that mammalian miRNAs are involved in regulation of target genes with a wide spectrum of molecular functions.

We also compared the proportion of genes that fell in each of the GO molecular function and GO biological process categories for the predicted targets of miRNAs, for targets of shuffled control sequences, and for the initial set of orthologous genes (Table 2 and Supplemental Table S4 on *Cell* website). The targets of the shuffled cohorts were enriched relative to the initial set of orthologous genes in certain GO biological process categories such as development (14% versus 8%) and transcription (13% versus 9%) (Table S4) and in molecular function categories such as nucleic acid binding (21% versus 14%), DNA binding (15% versus 10%), and transcriptional regulator activity (10% versus 6%) (Table 2). The biases seen for the shuffled cohorts are likely to result primarily from the TargetScan requirement for conserved segments in the 3′ UTRs of predicted targets and may reflect differences in the occurrence of 3′ UTR regulatory elements in different classes of genes.

In the GO biological process classifications, the predicted regulatory targets of authentic miRNA genes were enriched in the development category but no more than the targets of shuffled controls and were substantially more enriched for genes involved in transcription (21% of miRNA targets versus 13% of shuffled targets versus 9% of the initial dataset) and regulation of transcription (21% versus 12% versus 8%) (Supplemental Table S4). In terms of the GO molecular function classifications, targets of authentic miRNAs were enriched in the categories DNA binding (20% versus 15% versus

Table 2. Molecular Function Classification of Predicted miRNA Targets

GO ID	Molecular Function	miRNAs		Mean of Shuffled Cohorts		All Orthologous Genes	
	None/unknown	115	(29%)	45	(37%)	5131	(35%)
	Known function	285	(71%)	77	(63%)	9408	(65%)
GO:0005215	Transporter activity	36	(9%)	14	(12%)	1441	(10%)
GO:0005515	Protein binding	37	(9%)	11	(9%)	1005	(7%)
GO:0016787	Hydrolase activity	36	(9%)	12	(9%)	1502	(10%)
GO:0016740	Transferase activity	39	(10%)	10	(8%)	1104	(8%)
GO:0016301	Kinase activity	29	(7%)	6	(5%)	624	(4%)
GO:0046872	Metal ion binding	27	(7%)	5	(4%)	952	(7%)
GO:0003676	Nucleic acid binding	101	(25%)	26	(21%)	2072	(14%)
GO:0003677	DNA binding	80	(20%)	18	(15%)	1431	(10%)
GO:0030528	Transcription reg. act.	56	(14%)	12	(10%)	879	(6%)
GO:0000166	Nucleotide binding	52	(13%)	10	(8%)	1172	(8%)
GO:0004871	Signal transducer act.	55	(14%)	12	(10%)	1959	(13%)
GO:0004872	Receptor activity	29	(7%)	5	(4%)	1351	(9%)

The number and percentage of genes annotated with various Gene Ontology molecular function categories are shown for targets of nrMamm miRNAs, targets of shuffled control miRNAs (mean of four cohorts), and for the initial set of orthologous human-mouse-rat genes. If GO categories have a parent-child relationship, the child is indented. Because one gene can belong to multiple GO categories, the sum of the percentages in each column is not interpretable.

10%), transcription regulatory activity (14% versus 10% versus 6%), and nucleotide binding (13% versus 8% versus 8%) (Table 2). The differing numbers of predicted targets in the similar-sounding categories "regulation of transcription" (GO biological process classification) and "transcription regulatory activity" (GO molecular function classification) prompted us to investigate the gene content of these two categories. Inspection of the lists of genes showed that all but two of the predicted target genes in the "transcription regulatory activity" category were also included in the larger "regulation of transcription category," but that the latter category also contained more than two dozen additional target genes, the annotation of which generally supported a role in control of transcription. The GO process category "regulation of transcription" (Supplemental Table S4) therefore appears to provide a more complete listing of known and putative transcription factors.

The proportion of the predicted mammalian miRNA target genes involved in the GO process categories "transcription" and "regulation of transcription" was significantly higher than that seen for either shuffled targets or for the initial gene set ($p < 0.001$). Nonetheless, this bias was much lower in magnitude than that seen in plants: of the 49 targets predicted in a systematic search for complementarity to plant miRNAs, 69% were members of transcription factor gene families (Rhoades et al., 2002). Examples of other types of predicted mammalian targets include translational regulators (e.g., *COP9* subunit 6, *ERF1*), regulators of mRNA stability (e.g., *HU-Antigen D*), structural proteins (e.g., collagen), and enzymes (e.g., *G6PD*). The set of predicted miRNA targets conserved across all four vertebrates (Supplemental Table S5 online) was also somewhat biased toward genes involved in transcription, but had annotated functions consistent with the broad array of biological activities seen for the larger mammalian target set. We conclude that although mammalian miRNAs are sometimes at the center of gene regulatory networks, where they regulate genes, such as transcription factors, that regulate other genes, they are more likely than plant miRNAs to be at the periphery of the regulatory networks, where they regulate genes with a variety of molecular functions.

The predicted mammalian targets also differ from the plant targets with respect to biological function. Nearly all of the transcription factors (TFs) predicted to be plant miRNA targets have known or implied roles in plant development, as do several of the other predicted plant targets (Rhoades et al., 2002). By comparison, only ~13% of predicted mammalian miRNA targets were involved in development according to the GO biological process categories (Supplemental Table S4). An important caveat to this analysis is that gene annotation and GO categories are still evolving. Nonetheless, our data suggest that mammalian miRNAs are not exclusively, or even primarily, involved in the traditional miRNA role of developmental control. Instead, we find evidence for miRNA regulation of a very broad diversity of biological processes.

Experimental Procedures

MicroRNA Datasets

Human and mouse miRNA sequences that satisfy established criteria (Ambros et al., 2003a) were downloaded from the Rfam website (http://www.sanger.ac.uk/Software/Rfam). Human miRNAs that lacked annotated mouse orthologs and mouse miRNAs that lacked annotated human orthologs were searched against the mouse and human genomes, respectively, with BLASTN (Altschul et al., 1997) and MiRscan (Lim et al., 2003a, 2003b). To identify *Fugu* homologs, the human miRNAs were searched against the *Fugu* genome using BLASTN and MiRscan, and the 121 human miRNAs with perfectly homologous miRNAs in mouse and clear homologous miRNAs in *Fugu* were assigned to rMamm. For sets of human miRNAs in rMamm with identical seed heptamers, a single representative was chosen, yielding 79 human miRNAs (nrMamm). The choice was based on conservation to *Fugu* and *C. elegans* miRNAs when possible (i.e., the sequence most broadly conserved was chosen), but was otherwise essentially arbitrary (the miRNA with the lowest mir-# was generally chosen). The subset of 55 miRNAs from nrMamm that had perfect conservation to *Fugu* were assigned to nrVert.

3′ UTR Datasets

3′ UTR sequences for all human genes, and all mouse, rat, and *Fugu* genes associated with a human ortholog, were retrieved using

EnsMart version 15.1 (http://www.ensembl.org/EnsMart). Annotated 3′ UTR sequences were available for only 45% of rat genes in this set and for none of the *Fugu* genes. Moreover, 14% of annotated rat 3′ UTR sequences were less than 50 nucleotides in length. Therefore, we extended each annotated 3′ UTR with 2 kb of 3′ flanking sequence. Repetitive elements were masked in these sequences using RepeatMasker (Smit, A.F.A. and Green, P., http://repeatmasker.genome.washington.edu/cgi-bin/RM2_req.pl) with repeat libraries for primates, rodents, or vertebrates, as appropriate.

Identification of miRNA Target Sites

The 3′ UTR sequences were searched for antisense matches to the designated seed region of each miRNA (e.g., bases 2..8 starting from the 5′ end). Our choice of a 7 nt seed was motivated by the observation that shorter seeds gave substantially lower signal:noise ratios, while longer seeds reduced the number of predicted targets at comparable signal:noise ratios. Because changing the size of the seed has a large effect on the noise as well as the signal, these observations are much more difficult to interpret in terms of potential mechanistic implications than the "sliding seed" data of Figure 2B. For seeds located on the 5′ portion of the miRNA, 35 nt flanking the seed match on the 5′ end and 5 nt flanking the seed match on the 3′ end were retrieved (a "mirror" version of this algorithm was used for 3′ seeds in the experiment described in Figure 2B). Target sites in which the 35 nt flanking region contained masked bases or the seed match occurred less than 20 nt downstream of a previous seed match were discarded. Basepairing between the miRNA seed and UTR was extended with additional flanking basepairs as far as possible in both directions, allowing G:U pairs but disallowing gaps. The basepairing pattern of the remaining 3′ end (or in the case of a 3′ seed, the remaining 5′ end) was predicted by running RNAfold on a foldback sequence consisting of an artificial stemloop (5′-GGGCCCGGGULLLLLLACCCGGGCCC-3′, where "L" is an anonymous unpaired loop character, and all other bases are paired to a complementary base on the opposite side of the stem) attached to the extended seed match. RNAfold optimization was constrained so that all basepairs found in previous steps were fixed, the structure of the artificial stem was fixed, and bases in the miRNA and UTR were allowed to pair only with bases in the UTR and miRNA, respectively. The stemloop was removed, and RNAeval was used to estimate the energy of the miRNA:UTR duplex formed by the basepairs determined in the previous steps.

Parameter Optimization

Training sets were constructed with 40 randomly chosen miRNAs from nrMamm and 27 randomly chosen miRNAs from nrVert. The remaining microRNAs were assigned to the nrMamm and nrVert reference sets. TargetScan was tested on the training sets with various parameter values: T was varied from 5 to 25 in increments of 5, Z_C was varied between 1 and 10 in increments of 0.5, and R_C was varied between 50 and 1000 in increments of 50. The parameters $T = 20$, $Z_C = 4.5$, $R_C = 200$ were found to give an optimal signal:noise of 3.4:1 for the nrMamm training set. When R_C was raised to 300 or Z_C was lowered to 4, the signal:noise decreased only moderately to ~3:1. The parameters $T = 10$, $Z_C = 4.5$, $R_C = 350$ were found to give an optimal signal:noise of 4.6:1 for the nrVert training set used with UTR sets from all four genomes. For both the nrMamm and nrVert sets, the signal:noise ratios obtained using the training sets did not differ significantly from the corresponding signal:noise ratios obtained using the reference sets, and thus results from the two sets were merged.

Generation of Randomly Permuted Sequences

For each miRNA in nrMamm, randomly permuted sequences with the same starting base, length, and base composition as the real miRNA were generated until four sequences were found that deviated from the original miRNA by less than 15% in the following properties: (1) E(SM), the 1st order Markov probability of the seed match, (2) E(TM), the 1st order Markov probability of the antisense of the 3′ end of the miRNA (or the 5′ end in the case of a 3′ miRNA seed), (3) O(SM), the observed count of seed matches in the UTR dataset, and (4) the predicted folding free energy of a seed:seed match duplex. For a miRNA (or shuffled miRNA) with the initial sequence $S_1,S_2,S_3,S_4,S_5,S_6,S_7,S_8$, and the seed designated as bases 2..8, E(SM) was equal to $(P_{S_1,S_2} \cdot P_{S_2,S_3} \cdot P_{S_3,S_4} \cdot P_{S_4,S_5} \cdot P_{S_5,S_6} \cdot P_{S_6,S_7} \cdot P_{S_7,S_8})$ where $P_{S_k,S_{k+1}}$ was the conditional frequency of the nucleotide S_{k+1} given S_k at the previous position in the set of inverse complements of the UTRs in the UTR database. E(TM) was the analogous quantity calculated for the remainder of the sequence (i.e., for bases 9, 10, 11, … to the end of the miRNA or shuffled miRNA). O(SM) was determined directly from heptamer counts in the UTR dataset. The predicted folding free energy of a seed:seed match duplex was determined using RNAeval. The DiMirShuffle program generated shuffled controls for a given miRNA sequence by shuffling the dinucleotides of the specified miRNA seed (e.g., bases 2..8 of the miRNA).

DNA Constructs

The firefly luciferase vector was modified from pGL3 Control Vector (Promega), such that a short sequence containing multiple cloning sites (5′-AGCTCTATACGCGTCTCAAGCTTACTGCTAGCGT-3′) was inserted into the XbaI site immediately downstream from the stop codon. 3′UTR segments of the target genes were amplified by PCR from human genomic DNA and inserted into the modified pGL3 vector between SacI and NheI sites. PCR with the appropriate primers also generated inserts with point substitutions in the miRNA complementary sites. Wild-type and mutant inserts were confirmed by sequencing and are listed (Supplemental Figure S2 online).

Transfections and Assays

Adherent HeLa S3 cells were grown in 10% FBS in DMEM, supplemented with glutamine in the presence of antibiotics, to 80%–90% confluency in 24-well plates. Cells were transfected with 0.4 μg of the firefly luciferase reporter vector and 0.08 μg of the control vector containing *Renilla* luciferase, pRL-TK (Promega), in a final volume of 0.5 ml using Lipofectamine 2000 (Invitrogen). Firefly and *Renilla* luciferase activities were measured consecutively using the Dual-luciferase assays (Promega) 30 hr after transfection. Each firefly plasmid was tested in 12–15 transfections (four or five independent experiments, each with three culture replicates) involving two independent plasmid preparations (six to nine transfections each). A HeLa cell line that constitutively expressed miR-1 from a pol-II promoter was created using a derivative of the retroviral vector pRev-TRE (Clontech) containing a 500 bp fragment of human *mir-1d* gene. A HeLa S3 cell line that constitutively expressed miR-130 from the H1 pol-III promoter was constructed using a retroviral vector containing a 330 nt fragment of the human *mir-130* gene and a GFP gene under the murine 3-phosphoglycerate kinase promoter, which served as an infection marker (Chen, et al., 2003). Cells expressing GFP following infection were enriched to 95% purity by FACS.

Analysis of Gene Ontologies

Gene ontologies were assigned to human genes from the Ensembl database by crossreferencing Ensembl identifiers with GO identifiers using EnsMart version 15.1 (http://www.ensembl.org/EnsMart). The Gene Ontology Consortium database was retrieved from http://www.geneontology.org and function and process ontologies were compiled for all predicted target genes. In addition to the assigned categories, each gene was considered as having all more general ("parent") categories within the "Molecular Function" and "Biological Process" ontologies. In Tables 2 and S4, sets of GO categories were selected that were both broad enough to contain a significant fraction of the predicted targets and specific enough to be meaningful. Because the GO descriptions are not mutually exclusive, the sum of the percentages in these tables is not interpretable. GO categories were also used to produce the categories in Table 1. To be included in a category, a gene had to be annotated with at least one out of a set of GO categories. The sets of GO categories used were: regulation of transcription/DNA binding (GO:0003700, GO:0003713,GO:0003714, GO:0016563, or GO:0045449) and signal transduction/cell-cell signaling (GO:0004871, GO:0004872, GO:0007154, GO:0007165, GO:0007267 or GO:0008083).

Acknowledgments

We thank W.K. Johnston for technical assistance, C-Z. Chen and L.P. Lim for helpful discussions, H.F. Lodish for use of facilities

and equipment, N.C. Lau for the miR-1-expressing cell line, and G. Ruvkun for plasmids used to construct the *lin-41* reporters. Supported by grants from the N.I.H (D.P.B. and C.B.B.), the Searle Scholars Program (C.B.B.), and the Alexander and Margaret Stewart Trust (D.P.B.), and fellowships from the DOE (B.P.L.) and the Cancer Research Institute (I.S.).

Received: November 18, 2003
Revised: December 3, 2003
Accepted: December 4, 2003
Published: December 24, 2003

References

Abrahante, J.E., Daul, A.L., Li, M., Volk, M.L., Tennessen, J.M., Miller, E.A., and Rougvie, A.E. (2003). The *Caenorhabditis elegans* hunchback-like gene lin-57/hbl-1 controls developmental time and is regulated by microRNAs. Dev. Cell *4*, 625–637.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Ambros, V., Bartel, B., Bartel, D.P., Burge, C.B., Carrington, J.C., Chen, X., Dreyfuss, G., Eddy, S., Griffiths-Jones, S., Matzke, M., et al. (2003a). A uniform system for microRNA annotation. RNA *9*, 277–279.

Ambros, V., Lee, R.C., Lavanway, A., Williams, P.T., and Jewell, D. (2003b). MicroRNAs and other tiny endogenous RNAs in *C. elegans*. Curr. Biol. *13*, 807–818.

Aravin, A.A., Naumova, N.M., Tulin, A.A., Rozovsky, Y.M., and Gvozdev, V.A. (2001). Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in *Drosophila melanogaster* germline. Curr. Biol. *11*, 1017–1027.

Aukerman, M.J., and Sakai, H. (2003). Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. Plant Cell *15*, 2730–2741.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell, in press.

Brennecke, J., Hipfner, D.R., Stark, A., Russell, R.B., and Cohen, S.M. (2003). bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in *Drosophila*. Cell *113*, 25–36.

Chen, C.-Z., Li, L., Lodish, H.F., and Bartel, D.P. (2003). MicroRNAs modulate hematopoietic lineage differentiation. Science, in press. Published online December 4, 2003. 10.1126/science.1091903.

Chen, X. (2003). A MicroRNA as a translational repressor of APETALA2 in arabidopsis flower development. Science. Published online September 11, 2003. 10.1126/science.1088060.

Consortium, The Gene Ontology. (2001). Creating the gene ontology resource: design and implementation. Genome Res. *11*, 1425–1433.

Doench, J.G., Peterson, C.P., and Sharp, P.A. (2003). siRNAs can function as miRNAs. Genes Dev. *17*, 438–442.

Dostie, J., Mourelatos, Z., Yang, M., Sharma, A., and Dreyfuss, G. (2003). Numerous microRNPs in neuronal cells containing novel microRNAs. RNA *9*, 631–632.

Emery, J.F., Floyd, S.K., Alvarez, J., Eshed, Y., Hawker, N.P., Izhaki, A., Baum, S.F., and Bowman, J.L. (2003). Radial patterning of *Arabidopsis* shoots by class III HD-ZIP and KANADI genes. Curr. Biol. *13*, 1768–1774.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. Monatshefte fur Chemie *125*, 167–188.

Houbaviy, H.B., Murray, M.F., and Sharp, P.A. (2003). Embryonic stem cell-specific MicroRNAs. Dev. Cell *5*, 351–358.

Iseli, C., Stevenson, B.J., de Souza, S.J., Samaia, H.B., Camargo, A.A., Buetow, K.H., Strausberg, R.L., Simpson, A.J., Bucher, P., and Jongeneel, C.V. (2002). Long-range heterogeneity at the 3′ ends of human mRNAs. Genome Res. *12*, 1068–1074.

Kasschau, K.D., Xie, Z., Allen, E., Llave, C., Chapman, E.J., Krizan, K.A., and Carrington, J.C. (2003). P1/HC-Pro, a viral suppressor of RNA silencing, interferes with *Arabidopsis* development and miRNA function. Dev. Cell *4*, 205–217.

Krichevsky, A.M., King, K.S., Donahue, C.P., Khrapko, K., and Kosik, K.S. (2003). A microRNA array reveals extensive regulation of microRNAs during brain development. RNA *9*, 1274–1281.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. Science *294*, 853–858.

Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W., and Tuschl, T. (2002). Identification of tissue-specific microRNAs from mouse. Curr. Biol. *12*, 735–739.

Lagos-Quintana, M., Rauhut, R., Meyer, J., Borkhardt, A., and Tuschl, T. (2003). New microRNAs from mouse and human. RNA *9*, 175–179.

Lai, E.C. (2002). MicroRNAs are complementary to 3′UTR motifs that mediate negative post-transcriptional regulation. Nat. Genet. *30*, 363–364.

Lai, E.C. (2003). MicroRNAs: runts of the genome assert themselves Curr. Biol. *13*, R925–R936.

Lai, E.C., Tomancak, P., Williams, R.W., and Rubin, G.M. (2003). Computational identification of *Drosophila* microRNA genes. Genome Biol. *4:R42*, 1–20.

Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. (2001). An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. Science *294*, 858–862.

Lee, R.C., and Ambros, V. (2001). An extensive class of small RNAs in *Caenorhabditis elegans*. Science *294*, 862–864.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. Cell *75*, 843–854.

Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., and Bartel, D.P. (2003a). Vertebrate microRNA genes. Science *299*, 1540.

Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B., and Bartel, D.P. (2003b). The microRNAs of *Caenorhabditis elegans*. Genes Dev. *17*, 991–1008.

Lin, S.Y., Johnson, S.M., Abraham, M., Vella, M.C., Pasquinelli, A., Gamberi, C., Gottlieb, E., and Slack, F.J. (2003). The *C. elegans* hunchback homolog, hbl-1, controls temporal patterning and is a probable microRNA target. Dev. Cell *4*, p639–p650.

Llave, C., Xie, Z., Kasschau, K.D., and Carrington, J.C. (2002). Cleavage of scarecrow-like mRNA targets directed by a class of *Arabidopsis* miRNA. Science *297*, 2053–2056.

Moss, E.G., Lee, R.C., and Ambros, V. (1997). The cold shock domain protein LIN-28 controls developmental timing in *C. elegans* and is regulated by the lin-4 RNA. Cell *88*, 637–646.

Moss, E.G., and Tang, L. (2003). Conservation of the heterochronic regulator Lin-28, its developmental expression and microRNA complementary sites. Dev. Biol. *258*, 432–442.

Mourelatos, Z., Dostie, J., Paushkin, S., Sharma, A., Charroux, B., Abel, L., Rappsilber, J., Mann, M., and Dreyfuss, G. (2002). miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. Genes Dev. *16*, 720–728.

Nussinov, R. (1981). Nearest neighbor nucleotide patterns. Structural and biological implications. J. Biol. Chem. *256*, 8458–8462.

Palatnik, J.F., Allen, E., Wu, X., Schommer, C., Schwab, R., Carrington, J.C., and Weigel, D. (2003). Control of leaf morphogenesis by microRNAs. Nature *20*, 257–263. Published online August 20, 2003. 10.1038/nature01958.

Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M., Maller, B., Srinivasan, A., Fishman, M., Hayward, D., Ball, E., et al. (2000). Conservation across animal phylogeny of the sequence and temporal regulation of the 21 nucleotide *let-7* heterochronic regulatory RNA. Nature *408*, 86–89.

Reinhart, B.J., Slack, F.J., Basson, M., Bettinger, J.C., Pasquinelli, A.E., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. (2000). The 21 nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. Nature *403*, 901–906.

Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B., and Bartel, D.P. (2002). Prediction of plant microRNA targets. Cell *110*, 513–520.

Stark, A., Brennecke, J., Russell, R.B., and Cohen, S.M. (2003). Identification of *Drosophila* microRNA targets. PLOS Biol., in press. Published online October 13, 2003. 10.1371/journal.pbio.0000060.

Tang, G., Reinhart, B.J., Bartel, D.P., and Zamore, P.D. (2003). A biochemical framework for RNA silencing in plants. Genes Dev. *17*, 49–63.

Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. Cell *75*, 855–862.

Xu, P., Vernooy, S.Y., Guo, M., and Hay, B.A. (2003). The *Drosophila* MicroRNA Mir-14 suppresses cell death and is required for normal fat metabolism. Curr. Biol. *13*, 790–795.